Uniform Approximation of Linear Systems*

By HARRY HEFFES and PHILIP E. SARACHIK

(Manuscript received August 6, 1968)

A method for reducing the complexity of the class of linear, time-varying, dynamic control systems is developed where the approach taken is that of uniform approximation (that is, modeling for a region of initial conditions). The objective of the modeling procedure is to choose a linear time-invariant system of given dimension, that minimizes a "worst-case" type of error criterion. Some results from the theory of widths of sets in Banach space are used to obtain bounds on the optimal approximation error as a function of the dimension of the approximating system. The use of these bounds in choosing the order of the approximation is discussed. An example illustrates the use of the derived results.

I. INTRODUCTION

In the analysis and design of control systems it is often useful to have low order constant coefficient models for the system. The problem of modeling linear systems by lower order linear systems has received considerable attention, but these analyses have usually been restricted to the modeling of constant coefficient systems.

References 1 through 5 contain various approaches to the system approximation problem; however, these analyses are generally restricted to the modeling of constant coefficient systems or systems which are forced with a given input or initial condition.

The control system analyst often finds himself dealing with nonstationary systems, but little work has been done in the area of optimally modeling this class of systems. The emphasis here is on modeling the class of linear, homogeneous time-varying systems with constant coefficient models. Reference 6 considers approximation of forced systems. Rather than design the model requiring solutions of the actual and approximate systems be "close" for a prescribed initial

* From a dissertation written as part of the requirements for a Ph.D. degree, New York University, 1968.

condition, the approach taken here is that of uniform approximation. Initial conditions are assumed to lie in some set in Euclidean space and a "worst-case" type of error criterion is defined. This eliminates tuning the model to specific conditions which may not be met when using the model. The material presented here thus generalizes previous work in that it extends the class of systems considered to time-varying systems and generalizes the error criterion to handle the more realistic problem of modeling for regions of initial conditions.

The problem is of importance, for example, in trajectory analysis where the linear time-varying system is obtained by linearizing a set of nonlinear equations about a nominal trajectory. In this case the time-varying nature of the system is described by partial derivatives evaluated along the nominal trajectory. Solutions to the resulting equations require simulation for each set of initial conditions. Using a constant coefficient model eliminates the need for repeated simulation.

The above example illustrates the use of a simplified model in analysis. The designer is interested in reducing the complexity of high-order nonstationary control system plants since this provides a means for designing simpler controllers based upon the model description. The results presented here not only allow one to obtain stationary models but simultaneously offer the opportunity to obtain lower order models of the original system.

II. PROBLEM DEFINITION AND FORMULATION

The system we are considering is described by the linear, time-varying, homogeneous vector differential equation

$$\dot{x}(t) = A(t)x(t) \tag{1}$$

with the outputs given by

$$y(t) = C(t)x(t) \tag{2}$$

where

$x(t)$ is an n-vector

$A(t)$ is an $n \times n$ matrix whose elements are bounded and piecewise continuous on $[t_o , t_f]$.

$C(t)$ is an $m \times n$ matrix whose elements are bounded and piecewise continuous on $[t_o , t_f]$.

It is desired to obtain a constant coefficient system of $k$tb order* ($k \geqq m$)

$$\dot{\tilde{x}}(t) = \tilde{A}\tilde{x}(t) \tag{3}$$

such that the first m components of tbe state vector $\tilde{x}(t)$ closely approximate the components of $y(t)$ over the finite time interval $[t_o , t_f]$. Writing

$$\tilde{y}(t) = \tilde{C}\tilde{x}(t) \tag{4}$$

with

$$\tilde{C} = [I_{m\times m} \vdots 0]$$

the approximation problem can be viewed as cboosing the elements of the $k \times k$ matrix $\tilde{A}$ sucb that $\tilde{y}(t)$ approximates $y(t)$ over $[t_o , t_f]$.

Since, in general, it is not known at the time of modeling wbat initial conditions will exist in the system, it is desirable to have tbe approximating system depend on a prescribed range of initial conditions rather than being tuned to any specific initial condition. Tbe initial conditions are considered to lie in a closed, bounded convex subset of Euclidean n-space. That is,

$$x(t_o) \in R \subset E_n ,$$

and the performance criterion is given by

$$J_k(\tilde{A}) = \max_{x_o \in R} \min_{\tilde{x}_o \in E_k} \int_{t_o}^{t_f} (y - \tilde{y})' W(t)(y - \tilde{y})\, dt \tag{5}\dagger$$

where

$[t_o , t_f]$ is bounded
$y(t)$ is the solution of (1) and (2) with $x(t_o) = x_o$
$\tilde{y}(t)$ is the solution of (3) and (4) with $\tilde{x}(t_o) = \tilde{x}_o$
$W(t)$ is positive definite and bounded for all $t \in [t_o , t_f]$.

The above performance criterion corresponds to the worst case error in tbe approximation, corresponding to a given model, wben the initial condition on the model, $\tilde{x}(t_o)$, is cbosen optimally in terms of the initial conditions on the actual system. The modeling objective is to choose $\tilde{A}$ to minimize $J_k(\tilde{A})$(that is, minimize tbe maximum approximation error).

The approximation problem will be cast into a Hilbert space setting

* Notice that k is not restricted from above. It may be desirable to have $k > n$ if the original system is time-varying.
† In all that follows the prime denotes transpose.

which will permit the use of many of the general results to be presented in the next section. Vector spaces of solutions of the original system equations and any member of the class of approximate system equations are established. These spaces are then imbedded into an encompassing Hilbert space. It then is shown that the problem of finding an optimal approximation can be viewed as a problem of finding the "best" subspace (of a given form) of the Hilbert space to use in approximating solutions of the original system. Writing the output vector of the original system in terms of the transition matrix leads to

$$y(t) = C(t)\Phi(t, t_o)x(t_o) \tag{6}$$

where the transition matrix $\Phi(t, t_o)$ satisfies

$$\frac{d}{dt}\Phi(t, t_o) = A(t)\Phi(t, t_o) \tag{7}$$

with initial conditions

$$\Phi(t_o , t_o) = I. \tag{8}$$

Now if the original system is completely observable[8,9] on the finite interval $[t_o , t_f]$ the columns of the $m \times n$ matrix $C(t)\Phi(t, t_o)$ are linearly independent as vector-valued time functions. That is,

$$C(t)\Phi(t, t_o)x(t_o) \equiv 0 \quad \text{for all} \quad t \in [t_o , t_f]$$

implies $x(t_o) = 0$. For an observable system, the initial state can be determined uniquely from knowledge of the output. Since $x(t_o) = 0 \Rightarrow y(t) \equiv 0$ and, from observability, $y(t) \equiv 0 \Rightarrow x(t_o) = 0$ the linear independence of the columns of $C(t)\Phi(t, t_o)$ follows.

Let $\bar{\mathcal{Y}}$ be the linear space spanned by the n columns of $C(t)\Phi(t, t_o)$. The solutions of the original system lie in $\bar{\mathcal{Y}}$, which is of dimension n for an observable system. Notice that the number of components (m) in the vector y and the dimension of the space $\bar{\mathcal{Y}}$ need not be the same. If the system is not completely observable on $[t_o , t_f]$ the dimension of $\bar{\mathcal{Y}}$ is less than n.

The solutions to equations (3) and (4) can be written as

$$\tilde{y}(t) = \tilde{C}e^{\tilde{A}(t-t_o)}\tilde{x}(t_o) \tag{9}$$

where

$$e^{\tilde{A}(t-t_o)} = \sum_{i=0}^{\infty} \frac{\tilde{A}^i(t - t_o)^i}{i!}$$

and

$$\frac{d}{dt} e^{\bar{A}(t-t_0)} = \bar{A} e^{\bar{A}(t-t_0)}. \tag{10}$$

It is thus seen that solutions $\bar{y}(t)$ lie in the vector space spanned by the k columns of the $m \times k$ matrix $\bar{C}e^{\bar{A}(t-t_0)}$. Denote this vector space as $\mathcal{Y}_k$. If $\bar{A}$ is such that the approximate system is observable then $\mathcal{Y}_k$ is of dimension k and the k columns of $\bar{C}e^{\bar{A}(t-t_0)}$ form a basis

$$\{g_i ; i = 1, \cdots, k\}$$

for the k-dimensional vector space $\mathcal{Y}_k$ of approximating solutions. These basis elements can be written as

$$g_i(t) = \bar{C}e^{\bar{A}(t-t_0)}K_i \tag{11}$$

$$\mathbf{g}_i = \{g_i(t); t \in [t_0, t_f]\} \tag{12}$$

where K_i is the ith column of the $k \times k$ identity matrix. If the approximation is not observable the dimension of $\mathcal{Y}_k$ is less than k. In any case vector spaces $\mathcal{Y}_k$ with basis elements of the form (11) characterize the approximating systems where $\bar{A}$ is a $k \times k$ real matrix. Defining

$$\mathcal{D}_k = \{\mathcal{Y}_k ; \mathbf{g}_1, \cdots, \mathbf{g}_k \text{ span } \mathcal{Y}_k\} \tag{13}$$

where $g_i(t)$ is given by equation (11) and $\bar{A}$ is any real constant $k \times k$ matrix casts the problem into finding an element of $\mathcal{D}_k$ minimizing J_k.

The problem of finding an optimal approximation has been cast into the problem of finding an extremal space $\mathcal{Y}_k^* \in \mathcal{D}_k$ of approximating solutions. A Hilbert space $\mathcal{H}$ containing $\bar{\mathcal{Y}}$ and all members of $\mathcal{D}_k$ will now be constructed.

Recall that the elements of $\bar{\mathcal{Y}}$ and $\mathcal{Y}_k$ are real, vector-valued, time functions having m components. Thus each element of the Hilbert space $\mathcal{H}$ to be constructed will have m components. The inner product in $\mathcal{H}$ is defined by

$$(\mathbf{y}_1, \mathbf{y}_2) = \int_{t_0}^{t_f} y_1'(t) W(t) y_2(t)\, dt \tag{14}$$

where $W(t)$ is a real symmetric $m \times m$ matrix which is positive definite for $t \in [t_0, t_f]$ and whose elements are bounded for $t \in [t_0, t_f]$. Notice that this is the same matrix appearing in the performance criterion given by equation (5). The norm of an element in $\mathcal{H}$ is given by

$$||\mathbf{y}|| = (\mathbf{y}, \mathbf{y})^{\frac{1}{2}}. \tag{15}$$

The Hilbert space $\mathcal{H}$ is defined as

$$\mathcal{H} = \{\mathbf{y}; \mathbf{y} \text{ has } m \text{ components}, \|\mathbf{y}\| < \infty\}$$

where $\| \mathbf{y} \|$ is given by (15) and the inner product given by (14).

Since

$$t_f - t_o < \infty$$

and tbe elements of $A(t)$ and $C(t)$ are bounded it follows that solutions of equations (1) and (2) are bounded tbus yielding

$$\bar{\mathcal{Y}} \subset \mathcal{H}.$$

Since elements of $\mathcal{Y}_k$ are bounded over the finite interval $[t_o , t_f]$

$$\mathcal{Y}_k \subset \mathcal{H}.$$

That $\mathcal{Y}_k$ and $\bar{\mathcal{Y}}$ are subspaces of $\mathcal{H}$ follows from the fact tbat any finite-dimensional linear set in a normed space is closed[10].

The set of functions to be approximated are solutions to the original system equations with the initial conditions $x(t_o)$ satisfying

$$x(t_o) \in R \subset E_n$$

where R is a closed, bounded convex subset of Euclidean n-space. Writing

$$\mathcal{F} = \{\mathbf{y}; y(t) = C(t)\Phi(t, t_o)x(t_o), x(t_o) \in R\} \tag{16}$$

gives

$$J_k(\tilde{A}) = \max_{\mathbf{y} \in \mathcal{F}} \min_{\tilde{\mathbf{y}} \in \mathcal{Y}_k} \| \mathbf{y} - \tilde{\mathbf{y}} \|^2 \tag{17}$$

where the modeling objective is to find

$$\bar{d}_k^2 \triangleq \inf_{\mathcal{Y}_k \in \mathfrak{D}_k} \max_{\mathbf{y} \in \mathcal{F}} \min_{\tilde{\mathbf{y}} \in \mathcal{Y}_k} \| \mathbf{y} - \tilde{\mathbf{y}} \|^2. \tag{18}$$

Before proceding to solve the formulated approximation problem, some results from the theory of widths in Banach space are outlined. Lower bounds on the optimal performance are found as a function of the dimension of the approximating system.

III. WIDTHS OF SETS IN BANACH SPACE AND LOWER BOUNDS*

Classically, approximation theory was concerned with the following problem. Given a function to approximate and a set of approximating functions (sinusoids, exponentials, and polynomials, for example) find that linear combination of approximating functions which

* Ref. 7 contains an excellent treatment of widths of sets in Banach space.

minimizes some distance function. Notice tbat here the approximating functions are given as part of the prohlem statement.

Rather than approximate a single function, the problem under consideration is to approximate the class of functions $\mathfrak{F}$ given hy (16). For a given class of functions $\mathfrak{F}$ it is desired to obtain a "best" set of approximating functions rather than to choose the set arbitrarily. A measure of comparison is introduced which enables one to evaluate the efficiency of different sets of approximating functions. Tbe following definitions serve to illustrate these ideas.

Let $\mathfrak{B}$ be a Banach space containing a set of functions $\mathfrak{F}$ to be approximated by elements of an n-dimensional suhspace, X_n, of $\mathfrak{B}$. It is desired to find the "best" n-dimensional suhspace, or equivalently tbe "hest" set of approximating functions to use in approximating elements of $\mathfrak{F}$.

For a given $f \in \mathfrak{F}$ and $X_n \subset \mathfrak{B}$

$$\inf_{x \in X_n} \| f - x \|$$

represents how well one can do in approximating a given f with elements of X_n. Taking the supremum of the above quantity over all elements in $\mathfrak{F}$ leads to the following definition.

Definition 1: The *deviation* of $\mathfrak{F}$ from X_n is given by

$$E_{X_n}(\mathfrak{F}) = \sup_{f \in \mathfrak{F}} \inf_{x \in X_n} \| f - x \|.$$

The deviation represents the worst case approximation error over the class $\mathfrak{F}$ when using elements of X_n. Notice that the deviation serves as a performance measure of X_n. Taking the infimum of the deviation over all n-dimensional suhspaces of $\mathfrak{B}$ leads to the following definition.

Definition 2: The nth *width* of $\mathfrak{F}$ is given by

$$\begin{aligned} d_n(\mathfrak{F}) &= \inf_{X_n \subset \mathfrak{B}} E_{X_n}(\mathfrak{F}) \\ &= \inf_{X_n \subset \mathfrak{B}} \sup_{f \in \mathfrak{F}} \inf_{x \in X_n} \| f - x \|. \end{aligned}$$

Some of the elementary results following from the above definitions are

(*i*) The monotonicity of the width:

$$d_0(\mathfrak{F}) \geqq d_1(\mathfrak{F}) \geqq d_2(\mathfrak{F}) \geqq \cdots$$

and

(*ii*) Tbe nested property: If $\mathfrak{F}_1 \subset \mathfrak{F}_2 \subset \cdots$ then

$$d_n(\mathfrak{F}_1) \leqq d_n(\mathfrak{F}_2) \leqq \cdots .$$

Notice that

$$J_k(\tilde{A}) = E^2_{y_k}(\mathfrak{F}). \tag{19}$$

In defining $\tilde{d}_k^2$ the infimum of the square of the deviation was taken over the j-dimensional ($j \leqq k$) subspaces in $\mathfrak{D}_k$ whereas in defining d_k the infimum was taken over all k-dimensional subspaces of $\mathfrak{G}$. Using the monotonicity property of the width, with $\mathfrak{K}$ serving as the required Banach space, gives

$$J_k(\tilde{A}) \geqq d_k^2(\mathfrak{F}) \tag{20}$$

for any $k \times k$ matrix $\tilde{A}$.

Definition 3: U_n is a *closed ball of radius* r in X_n if

$$U_n = \{x \in X_n ; \| x \| \leqq r\}.$$

The following theorem, by Gohberg and Krein, is proved in Ref. 7 and will he found useful.

Theorem: *If* X_{n+1} *is an* $(n + 1)$*-dimensional subspace of a Banach space* $\mathfrak{G}$ *and if* U_{n+1} *is the closed ball of radius* r *in* X_{n+1} *then* $d_n(U_{n+1}) = r$.

This theorem and the nested property of widths can he used to obtain lower bounds on $d_n(\mathfrak{F})$. This lower hound can he obtained by constructing a ball in an $(n + 1)$-dimensional subspace and choosing r such that $U_{n+1} \subset \mathfrak{F}$. Using the nested property then leads to

$$r = d_n(U_{n+1}) \leqq d_n(\mathfrak{F}). \tag{21}$$

Since

$$\tilde{d}_k \geqq d_k \tag{22}$$

the radius of ball also serves as a lower hound on $(J_k)^{\frac{1}{2}}$.

Lemma 1: *Let* $\Phi(t, t_o)$ *and* $C(t)$ *be the transition matrix and output matrix, respectively, of the original system* (1) *and* (2). *Assume this system to be completely observable on* $[t_o , t_f]$. *Let* $W(t)$ *satisfy the previously stated conditions. Then the matrix*

$$M = \int_{t_o}^{t_f} \Phi'(t, t_o)C'(t)W(t)C(t)\Phi(t, t_o)\, dt \tag{23}$$

is positive definite.

Proof: Consider the quadratic form $x_0'Mx_0 = \| \mathbf{y} \|^2 \geqq 0$ where

$$x(t_o) = x_0 , \quad \text{that is,} \quad y(t) = C(t)\Phi(t, t_o)x_0 .$$

Now $\| \mathbf{y} \|^2 = 0 \Rightarrow y(t) \equiv 0$ on $[t_o , t_f]$.

Since the system is observable $y \equiv 0 \Rightarrow x_0 = 0$. Thus M is positive definite.

Tbe following theorem provides tbe lower bound on the performance.

Theorem 1: *Let R be the closed region of initial conditions on the original system and let $x(t_o) = 0$ be an interior point of R. Assume the system to be completely observable on $[t_o, t_f]$. Denote the boundary of R by ∂R and let*

$$\rho^2 \triangleq \min_{x(t_o) \in \partial R} x'(t_o)x(t_o). \tag{24}$$

Let the eigenvalues of the positive definite matrix M be ordered $\lambda_1(M) \geqq \lambda_2(M) \geqq \cdots \geqq \lambda_n(M)$. Then the performance, for any k-dimensional approximating system, satisfies $J_k(\tilde{A}) \geqq \rho^2\lambda_{k+1}(M)$ for $k < n$.

Proof: Let

$$\mathfrak{F} = \{\mathbf{y};\ y(t) = C(t)\Phi(t, t_o)x(t_o),\ x(t_o) \in R\}.$$

A $k + 1$ dimensional ball will now be constructed which is a subset of $\mathfrak{F}$. Consider the $k + 1$ dimensional ball of radius r

$$U_{k+1} = \{\mathbf{y};\ y(t) = C(t)\Phi(t, t_o)x(t_o),\ x(t_o) \in E_{k+1} \subset E_n,\ \| \mathbf{y} \| \leqq r\}$$

E_{k+1} and r will be chosen such tbat $U_{k+1} \subset \mathfrak{F}$. Since M is real and symmetric it can be diagonalized with an orthogonal matrix T. Tbus $M = T'\Lambda T$ and

$$\| \mathbf{y} \|^2 = [Tx(t_o)]'\Lambda[Tx(t_o)] = z'\Lambda z$$

where

$$T' = T^{-1}$$

$$\Lambda = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix}$$

and

$$z = Tx(t_o).$$

Defining

$$E_{k+1} = \{x(t_o);\ [Tx(t_o)]_i = 0, \qquad i = k + 2, \cdots, n\}.$$

and

$$r^2 = \rho^2\lambda_{k+1}(M)$$

gives

$$U_{k+1} = \{\mathbf{y};\ y(t) = C(t)\Phi(t, t_o)x(t_o),$$
$$\|\mathbf{y}\|^2 \leqq \rho^2\lambda_{k+1}(M),\ \ [Tx(t_o)]_i = 0$$
$$i = k+2, \cdots, n\}.$$

Thus for $\mathbf{y} \in U_{k+1}$

$$\|\mathbf{y}\|^2 = x'(t_o)Mx(t_o) = \sum_{i=1}^{n} z_i^2\lambda_i \leqq \rho^2\lambda_{k+1}\ .$$

Since

$$z_i = 0 \qquad i = k+2, \cdots, n$$

and

$$\frac{\lambda_i}{\lambda_{k+1}} \geqq 1 \qquad \text{for} \quad i \leqq k+1$$

we have

$$\sum_{i=1}^{n} z_i^2 = x_0'x_0 \leqq \rho^2.$$

It then follows, from the definition of ρ^2 and the fact that zero is an interior point of R, that $x_0 \in R$ and therefore $\mathbf{y} \in \mathfrak{F}$. Thus $U_{k+1} \subset \mathfrak{F}$ and the desired result

$$J_k(\bar{A}) \geqq d_k^2(\mathfrak{F}) \geqq \rho^2\lambda_{k+1}(M) \qquad k < n$$

follows.

Remarks: Recalling that tbe eigenvalues of M are ordered, we notice that the lower bound is a decreasing function of the dimension of the approximating system. This result can be used to determine what order aproximating system (at least) need be considered to achieve a given performance. We emphasize that the bound depends on the original system and is obtainable prior to the modelling procedure. From an engineering viewpoint, if one has an approximating system whose performance is "close" to the bound it may not be necessary to seek the minor improvement. Notice that the only property of R appearing in the lower bound is ρ and no attempt was made to take the orientation of the set into account. The bound will therefore be least conservative when R is a hypersphere of radius ρ.

IV. EVALUATING THE PERFORMANCE FUNCTION

In this section the problem of finding the performance (or equivalently the deviation) of a given approximating system is considered. The optimal choice of initial conditions on the approximating system is obtained using some elementary Hilbert space concepts and it is shown that

$$\inf_{\bar{y} \in \mathcal{Y}_k} \| y - \bar{y} \|^2$$

is a positive semidefinite quadratic form in $x(t_0)$. Next, properties of convex functions are used to evaluate the performance for different classes of regions of initial conditions; namely, for ellipsoids and convex polyhedra. The Powell algorithm for minimizing a function of several variables, without calculating derivatives, is then outlined and applied to the system approximation problem.

The problem of finding

$$\delta^2 = \inf_{\bar{y} \in \mathcal{Y}_k} \| y - \bar{y} \|^2 \tag{25}$$

is equivalent to finding the best choice of initial conditions on a given approximating system characterized by $\mathcal{Y}_k \in \mathcal{D}_k$. It can be shown[11] that there exists a unique $\bar{y}^* \in \mathcal{Y}_k$ (y^* is called the projection of y in the space $\mathcal{Y}_k$) such that

$$\delta^2 = \| y - \bar{y}^* \|^2 = \| y \|^2 - \| \bar{y}^* \|^2. \tag{26}$$

Furthermore, since $g_1, g_2, \cdots, g_k$ spans $\mathcal{Y}_k$, $\bar{y}^*$ has the representation

$$\bar{y}^* = \sum_{i=1}^{k} g_i \bar{x}_i^*$$

where

$$G(g_1, \cdots, g_k)\bar{x}^* = \begin{bmatrix} (y, g_1) \\ \vdots \\ (y, g_k) \end{bmatrix} \tag{27}$$

$$\bar{x}^* = \begin{bmatrix} \bar{x}_1^* \\ \vdots \\ \bar{x}_k^* \end{bmatrix}$$

and G is the Grammian of $\{\mathbf{g}_i ; i = 1, \cdots, k\}$, that is,

$$[G(\mathbf{g}_1, \cdots, \mathbf{g}_k)]_{i,j} = (\mathbf{g}_i, \mathbf{g}_j) \qquad i, j = 1, \cdots, k.$$

Any solution to (27) results in an optimum choice of initial conditions on the approximate system. If the $\mathbf{g}_i$'s are linearly independent (this corresponds to the system being observable) the Grammian is invertible and $\hat{x}^*$ is unique. Thus

$$\hat{x}^*(t_0) = G^\dagger(\mathbf{g}_1, \cdots, \mathbf{g}_k)\begin{bmatrix}(\mathbf{y}, \mathbf{g}_1)\\ \vdots \\ (\mathbf{y}, \mathbf{g}_k)\end{bmatrix} \tag{28}$$

where $G^\dagger$ is the pseudoinverse[12] of G.

The Grammian is given by

$$G(\mathbf{g}_1, \cdots, \mathbf{g}_k) = \int_{t_0}^{t_f} e^{\tilde{A}'(t-t_0)}\tilde{C}'W(t)\tilde{C}e^{\tilde{A}(t-t_0)}\, dt \tag{29}$$

and

$$(\mathbf{y}, \mathbf{g}_i) = K_i'Fx(t_0) \tag{30}$$

where F is given by

$$F = \int_{t_0}^{t_f} e^{\tilde{A}'(t-t_0)}\tilde{C}'W(t)C(t)\Phi(t, t_0)\, dt. \tag{31}$$

Using (30) in (28) gives

$$\hat{x}^*(t_0) = G^\dagger Fx(t_0). \tag{32}$$

Thus the optimal initial condition on the approximating system is obtained by linearly transforming the actual initial condition with the $(k \times n)$ matrix $G^\dagger F$. Using the ortbogonality property (26) yields

$$\| \mathbf{y} - \tilde{\mathbf{y}}^* \|^2 = \| \mathbf{y} \|^2 - \hat{x}^{*\prime}(t_0)G\hat{x}^*(t_0).$$

Letting

$$M = \int_{t_0}^{t_f} \Phi'(t, t_0)C'(t)W(t)C(t)\Phi(t, t_0)\, dt \tag{33}$$

and using (32) and the symmetry of G (and thus $G^\dagger$) gives

$$\| \mathbf{y} - \tilde{\mathbf{y}}^* \|^2 = x'(t_0)(M - F'G^\dagger F)x(t_0). \tag{34}$$

In summary,

$$\delta^2 = \inf_{\tilde{\mathbf{y}} \in \mathcal{Y}_k} \| \mathbf{y} - \tilde{\mathbf{y}} \|^2 = x'(t_0)\, Dx(t_0) \tag{35}$$

with

$$D = M - F'G^{\dagger}F. \tag{36}$$

Thus, finding the optimal initial condition on the approximating systems leads to the positive semidefinite quadratic form (35) for the approximation error. The above represents the first step in evaluating the performance of any given approximating system.

Since D is a positive semidefinite matrix, δ^2 defined by (35) is a convex function of the initial state $x(t_o)$. The following theorem from Ref.13 is useful in maximizing δ^2.

Theorem: If the absolute maximum of a convex function, defined on a closed, bounded, convex set, is finite then the absolute maximum is taken on at an extreme point of the set.

Remarks: An extreme point of a convex set is a point in the set that cannot be written as a convex combination of two other points in the set. Notice that an extreme point is a boundary point; however, generally not every boundary point is an extreme point. Thus, if one is seeking the absolute maximum of a convex function defined on a closed, bounded, convex set only boundary points need be considered. Also if the domain of definition is a convex polyhedron (a closed, bounded, convex set with a finite number of extreme points) the absolute maximum can be obtained by simply evaluating the function at the extreme points and choosing the largest value.

Two general classes of closed, bounded, convex regions of initial conditions are considered in this paper, the ellipsoid and the convex polyhedron.

Let the region under consideration be an ellipsoid defined by

$$R = \{x(t_o);\quad x'(t_o)Bx(t_o) \leqq r^2\} \tag{37}$$

where B is a positive definite, symmetric matrix and r is finite. Notice that R is closed, bounded, and convex. Now the constrained maximization problem is one with an inequality constraint. Using the convexity of R and δ^2, the absolute maximum of tbe quadratic form is seen to take place on the boundary of the set R. Thus the performance can be written

$$J_k(\bar{A}) = \max_{x(t_o)} x'(t_o)\ Dx(t_o)$$

subject to the constraint

$$x'(t_o)Bx(t_o) = r^2.$$

It can easily be shown that the $x(t_0)$ maximizing the quadratic form is the eigenvector of the matrix $B^{-1}D$ corresponding to the largest eigenvalue and the maximum is given by

$$J_k(\tilde{A}) = \lambda_{\max}(B^{-1}D)r^2. \tag{38}$$

A convex polyhedron is usually representative of the type of information one has as to the range of initial conditions. As an example of this situation consider the original system to represent linearized equations of motion of a space vehicle. Suppose it is known that the range of initial conditions are in terms of bounds on position, velocity deviations, and so on. For example,

$$| x_1(t_0) | \leqq 100 \text{ feet.}$$

$$| x_2(t_0) | \leqq 5 \text{ feet per second.}$$

This particular region is described by a rectangular region in state space with the extreme points being the corners

$$\begin{bmatrix} 100 \\ 5 \end{bmatrix}, \quad \begin{bmatrix} 100 \\ -5 \end{bmatrix}, \quad \begin{bmatrix} -100 \\ 5 \end{bmatrix}, \quad \begin{bmatrix} -100 \\ -5 \end{bmatrix}.$$

In general for this type of initial condition region, that is,

$$| x_i(t_0) | \leqq b_i \qquad i = 1, \cdots, n,$$

the region has 2^n extreme points. Since δ^2 is an even function of $x(t_0)$ it is only necessary to consider 2^{n-1} extreme points eliminating from consideration the negative of any point considered.

The convex polyhedron region also is important, for example, since it may be used to simply approximate a more complex region. In general, let

$$x^{(i)} \; i = 1, 2, \cdots, N$$

be the extreme points of the convex polyhedron R. Using the convexity of δ^2 in the initial state $x(t_0)$ the absolute maximum δ^2 over R takes place at one of the $x^{(i)}$. Letting

$$\delta_i^2 = x^{(i)\prime} D x^{(i)}$$

where D is given by equation (36) leads to

$$J_k(\tilde{A}) = \max [\delta_1^2, \delta_2^2, \cdots, \delta_N^2]. \tag{39}$$

V. MINIMIZING THE PERFORMANCE FUNCTION

Since it is a fairly simple matter to evaluate the performance, whereas evaluating the gradient of the performance function requires significant computational effort, it is desirable to use a numerical procedure not requiring a gradient computation. Notice that $J_k(\tilde{A})$ is not generally differentiable. Here, for completeness, the Powell method of minimizing a function of several variables without calculating derivatives is presented.[14] Reference 15 contains a summary of the various minimization techniques available not requiring the computation of a derivative. See Refs. 14 and 15 for a more detailed description of the methods and their convergence properties.

Consider a real, scalar, valued function of N real variables $a_1, \cdots, a_N$ written $f(a)$. Powell's iterative scheme concerns itself with finding the minimum of $f(a)$ without computing its derivative.

Each iteration of the modified Powell procedure starts with a search down N linearly independent directions

$$\eta_1, \eta_2, \cdots, \eta_N$$

starting with an initial guess a_o and defines a new set of directions for the next iteration.

An iteration of the recommended procedure, suggested by Powell, is:

(*i*) for $j = 1, 2, \cdots, N$ calculate λ_j such that $f(a_{j-1} + \lambda_j \eta_j)$ is minimum and define $a_j = a_{j-1} + \lambda_j \eta_j$.

(*ii*) Find the integer m, $1 \leqq m \leqq N$, such that $f(a_{m-1}) - f(a_m)$ is a maximum and define $\Delta = f(a_{m-1}) - f(a_m)$.

(*iii*) Calculate $f_3 = f(2a_N - a_o)$ and define

$$f_1 = f(a_o)$$

$$f_2 = f(a_N).$$

(*iv*) If either $f_3 \geqq f_1$ or

$$(f_1 - 2f_2 + f_3)(f_1 - f_2 - \Delta)^2 \geqq \tfrac{1}{2}\Delta(f_1 - f_3)^2$$

use the old directions $\eta_1, \cdots, \eta_N$ for the next iteration and use a_N for the next a_o, otherwise

(*v*) define $\eta = a_n - a_o$ and calculate λ such that $f(a_N + \lambda\eta)$ is minimum. Use

$$\eta_1, \cdots, \eta_{m-1}, \eta, \eta_{m+1}, \cdots, \eta_N$$

as the new directions and $a_N + \lambda\eta$ as the new a_o.

The performance functions, for the two classes of initial conditions being considered are given by (38) and (39) in terms of the matrix D defined in (36). The major effort in computing the performance function is seen to lie in the computation of D. Sylvester's expansion (see page 83 of Ref. 16) for computing $e^{\tilde{A}t}$ is useful in the computation of the matrices F and G.

The basic procedure can be outlined as follows:

(*i*) Compute and store $C(t)\Phi(t, t_o)$ for $t \in [t_o, t_f]$ using (7) and (8).

(*ii*) Evaluate M using (23).

(*iii*) If it is desired to compute the lower bounds to aid in choosing the dimension of the approximating system, compute the eigenvalues of M and obtain the bounds from the result of Theorem 1.

(*iv*) Choose starting values for $\tilde{A}$ and choose the directions for the initial search in the modified Powell method to be

$$\begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \cdots, \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}$$

where the above are k^2 vectors.

(*v*) Use modified Powell method to determine the minimum of the performance function. Each element of the vector a in the Powell method corresponds to an element of $\tilde{A}$.

VI. EXAMPLE

A linearized missile guidance loop may be expressed in the form

$$\dot{x}_1 = x_2 , \qquad \dot{x}_2 = \frac{H}{m - t} x_3 , \qquad \dot{x}_3 = u \tag{40}$$

where x_1 is the lateral position deviation from a nominal trajectory, x_2 is the lateral velocity deviation, x_3 is the attitude deviation in the given direction and u is the control signal. The relationship between the attitude and lateral acceleration is given through the time-varying gain $H/(m - t)$ which accounts for the loss of mass because of fuel consumption.

Suppose it is desired to approximate homogeneous solutions to (40) for initial conditions (at beginning of a stage) lying in a set R (R is defined later) with solutions of a constant coefficient system. The actual system (40) can be written in the vector-matrix form

$$\dot{x}(t) = A(t)x(t) \tag{41}$$

with output

$$y(t) = [1\ 0\ 0]x(t) = Cx(t) \tag{42}$$

where

$$x(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \end{bmatrix}$$

and

$$A(t) = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & \dfrac{H}{m - t} \\ 0 & 0 & 0 \end{bmatrix}. \tag{43}$$

Let

$$|| y ||^2 = \int_0^T y^2(t)\, dt. \tag{44}$$

Before proceeding to find the approximation it is instructive to determine the lower bounds on the optimal performance. This will naturally aid in choosing the dimension of the approximating system. The matrix, M, defined by (33), is given by

$$M = \int_0^T \Phi'(t, o)C'C\Phi(t, o)\, dt \tag{45}$$

with

$$\frac{d}{dt}\Phi(t, o) = A(t)\Phi(t, o). \tag{46}$$

The transition matrix, which is the solution to (46) with the identity initial condition, is given by

$$\Phi(t, o) = \begin{bmatrix} 1 & t & H\left\{(m - t) \ln \left(\frac{m - t}{m}\right) + t\right\} \\ 0 & 1 & -H \ln \left(\frac{m - t}{m}\right) \\ 0 & 0 & 1 \end{bmatrix}.$$

Evaluating M leads to

$$M = \begin{bmatrix} T & \frac{T^2}{2} & H\left\{\frac{T^2}{2} - \frac{(m - T)^2}{2} \ln \left(\frac{m - T}{m}\right) + \frac{1}{4}(T^2 - 2mT)\right\} \\ \frac{T^2}{2} & \frac{T^3}{3} & M_{23} \\ M_{13} & M_{23} & M_{33} \end{bmatrix}$$

with

$$M_{23} = H\left[\frac{T^3}{3} - \frac{5}{36} m^3 - \frac{(2T + m)(m - T)^2}{6} \cdot \ln \left(\frac{m - T}{m}\right) + \frac{(m - T)^2(4T + 5m)}{36}\right]$$

and

$$M_{33} = H^2\left[\frac{T^3}{3} - \frac{(m - T)^3}{3} \ln^2 \left(\frac{m - T}{m}\right) + \frac{2}{9}(m - T)^3 \cdot \ln \left(\frac{m - T}{m}\right) + \frac{2}{27}\{m^3 - (m - T)^3\} - \frac{10}{36} m^3 - \frac{(2T + m)}{3}(m - T)^2 \cdot \ln \left(\frac{m - T}{m}\right) + \frac{(m - T)^2(4T + 5m)}{18}\right].$$

Let the constants defining the problem be given by

$m = 15$ seconds (normalized mass)

$T = 10$ seconds

$H = 15$ (pound-seconds per slug $) \times 10^{-3}$

and let the region of initial conditions be given by

$$R = \{x(o);\ |x_1(o)| \leqq 30 \text{ feet},\ |x_2(o)| \leqq 2 \text{ feet per second},\ |x_3(o)| \leqq 1 \text{ milliradian}\}.$$

Evaluating M for the above values of the constants leads to

$$M = \begin{bmatrix} 10 & 50 & 206 \\ 50 & 333 & 1570 \\ 206 & 1570 & 8082 \end{bmatrix}$$

with eigenvalues

$$\lambda_1 = 8393, \qquad \lambda_2 = 31, \qquad \lambda_3 = 1.1.$$

We have

$$J_0 \geqq 8{,}393$$

$$J_1 \geqq 31$$

and

$$J_2 \geqq 1.1.$$

Here J_0 represents

$$\max_{x_0 \in R} \| \mathbf{y} \|^2.$$

The second order approximation thus has the possibility of yielding a negligible approximation error. Thus in the remainder of this paper the optimal second order approximation will be sought. Thus

$$\dot{\tilde{x}} = \tilde{A}\tilde{x} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \tilde{x}$$

and

$$\tilde{y} = [1 \ 0]\tilde{x}.$$

Tbe initial choice for $\tilde{A}$ in the iterative procedure is

$$\tilde{A}_0 = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$$

which represents polynomial approximations to solutions of the original system.

The extreme points of R are given by

$$x^{(1)} = \begin{bmatrix} 30 \\ 2 \\ 1 \end{bmatrix}, \quad x^{(2)} = \begin{bmatrix} -30 \\ 2 \\ 1 \end{bmatrix}, \quad x^{(3)} = \begin{bmatrix} 30 \\ -2 \\ 1 \end{bmatrix} \quad \text{and} \quad x^{(4)} = \begin{bmatrix} 30 \\ 2 \\ -1 \end{bmatrix}$$

and their negatives. Thus

$$J_2(\tilde{A}_o) = \max_i \{x^{(i)\prime} Dx^{(i)}\} = 340$$

where D is evaluated from (36). It is thus seen that the performance function is far greater than the lower bound and the possibility exists for a significant improvement. The result of applying the Powell algorithm to this problem yields

$$\tilde{A}^* = \begin{bmatrix} 0.244 & 0.827 \\ 0.177 \times 10^{-3} & 0.629 \times 10^{-3} \end{bmatrix}$$

and

$$J_2(\tilde{A}^*) = 33.4$$

with the eigenvalues of $\tilde{A}^*$ given by

$$\lambda_1(\tilde{A}^*) = 0.245$$

$$\lambda_2(\tilde{A}^*) = 0.30 \times 10^{-4}.$$

The above results are obtained after three iterations of the Powell algorithm. The G, F and D matrices are given by

$$G = \begin{bmatrix} 271 & 771 \\ 771 & 2230 \end{bmatrix}$$

$$F = \begin{bmatrix} 43.14 & 296.0 & 1459 \\ 112.2 & 832.6 & 4241 \end{bmatrix}$$

and

$$D = \begin{bmatrix} 4.4 \times 10^{-6} & -1.8 \times 10^{-4} & 1.2 \times 10^{-4} \\ -1.8 \times 10^{-4} & 7.3 & 3.5 \times 10^{-3} \\ 1.2 \times 10^{-4} & 3.5 \times 10^{-3} & 4.2 \end{bmatrix}.$$

Evaluating

$$\max_i \{x^{(i)} Dx^{(i)}\}$$

gives the maximum approximation error occurring at the extreme point

$$x^{(3)} = \begin{bmatrix} 30 \\ -2 \\ 1 \end{bmatrix}.$$

Figure 1 shows the solution of the actual and approximate system for this worst-case initial condition. The solutions are obtained from

$$y(t) = 30 - 2t + \Phi_{13}(t, o)$$

and

$$\tilde{y}(t) = [1\ 0]e^{\tilde{A}*t}G^{-1}F\begin{bmatrix} 30 \\ -2 \\ 1 \end{bmatrix}.$$

$$\tilde{y}(t) = 4.82\, e^{\lambda_1 t} + 19.78\, e^{\lambda_2 t}.$$

The matrix relating the initial conditions is given by $G^{-1}F$, that is,

$$\tilde{x}(o) = G^{-1}Fx(o).$$

$$\tilde{x}(o) = \begin{bmatrix} 1.00 & 1.86 & -1.68 \\ -0.295 & -0.271 & 2.48 \end{bmatrix}x(o).$$

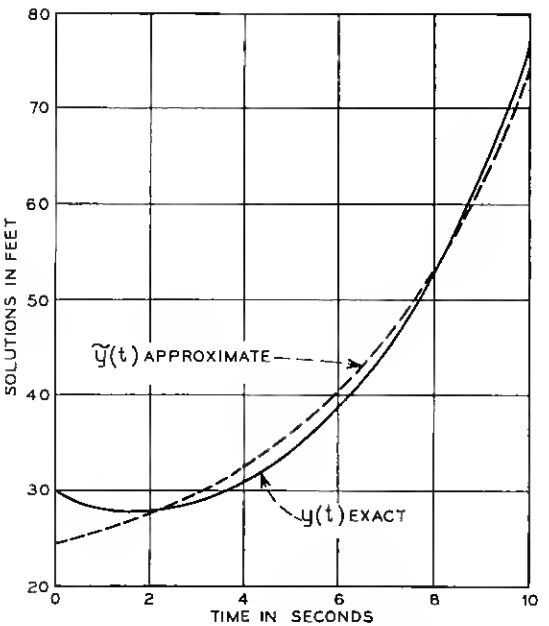

Fig. 1 — Exact and approximate solutions in worst case.

VII. CONCLUSIONS

A method for uniformly approximating solutions of linear, time-varying, homogeneous differential equations has heen presented. The problem of approximating systems subject to control or reference inputs is considered in Ref. 6 for the class of exponential polynomial control inputs.

One of the objectives of modeling with constant coefficient systems was to ohtain closed form approximations. Use of Sylvester's expansion allows one to derive these closed form expressions. However, more general classes of approximating systems can he sought while still maintaining the property that approximations are in closed form. For example, a general model of the form

$$\dot{\tilde{x}} = p(t)\tilde{A}\tilde{x}$$

where $p(t)$ is a scalar valued function possesses the closed form solution

$$\tilde{x}(t) = \exp\left[\tilde{A}\int_{t_0}^{t} p(\tau)\,d\tau\right]\tilde{x}(t_0)$$

and $p(t)$ as well as $\tilde{A}$ may he sought as part of the modeling procedure. A complexity constraint can be imposed on $p(t)$ by considering it to be a polynomial of given degree and the search for the model reduces to finding the coefficients of the polynomial as well as the elements of $\tilde{A}$.

VIII. ACKNOWLEDGMENTS

We greatly appreciate the valuahle suggestions of D. L. Jagerman.

Harry Heffes thanks Bell Telephone Laboratories, Incorporated, for their financial assistance under the doctoral support plan; Philip Sarachik thanks the Air Force Office of Scientific Research for their financial assistance under grant AF OSR 747-66.

REFERENCES

1. McBride, L. E., Jr., Schaefgen, H. W., and Steigler, K., "Time-Domain Approximation by Iterative Methods," IEEE Trans. Circuit Theory *CT-13*, No. 4, (December 1966), pp. 381–387.
2. Mitra, D., "On the Reduction of Complexity of Linear Dynamic Models," United Kingdom Atomic Energy Authority Report AEEW-R520, 1967.
3. Davison, E. J., "A Method for Simplifying Linear Dynamic Systems," IEEE Trans. Automatic Control, Vol. *AC-11*, No. 1, pp. 93–101, January 1966.
4. Nordahl, D. H., and Melsa, J. L., "Modeling with Lyapunov Functions," Proc. Joint Automatic Control Conf., 1967, pp. 208–215.
5. Meier, L., III, "Approximation of Linear Constant Systems by Linear

Constant Systems of Lower Order," Ph.D. Dissertation, Stanford, Calif.: Stanford University, 1965.
6. Heffes, H., "Approximation of Linear Time-Varying Systems by Linear Constant Coefficient Systems Over Finite-Time Intervals," Doctoral Dissertation, New York University, June 1968.
7. Lorentz, G. G., *Approximation of Functions,* New York: Holt, Reinhart and Winston, 1966.
8. Kreindler, E. and Sarachik, P. E., "On the Concepts of Controllability and Observability of Linear Systems," IEEE Tran. Automatic Control, *AC-9,* No. 2, (April 1964), pp. 129–136.
9. Kalman, R. E., "Mathematical Description of Linear Dynamical Systems," *J. SIAM Control, 1,* No. 2 (1963), pp. 152–192.
10. Kantorovich, L. V. and Akilov, G. P., *Functional Analysis in Normed Spaces,* New York: MacMillan, 1964, p. 58.
11. Akhiezer, N. I. and Glazman, I. M., *Theory of Linear Operators in Hilbert Space,* New York: Ungar, 1961.
12. Penrose, R., "A Generalized Inverse for Matrices," Proc. Cambridge Phil. Soc., *51,* pt. 3 (July 1955), pp. 406–413.
13. Hadley, G., *Nonlinear and Dynamic Programming,* New York: Addison Wesley, 1964.
14. Powell, M. J. D., "An Efficient Method of Finding the Minimum of a Function of Several Variables Without Calculating Derivatives," The Computer Journal, *7,* No. 2 (1964), pp. 155–162.
15. Fletcher, R., "Function Minimization Without Evaluating Derivatives—A Review," The Computer Journal, *8,* (1966), pp. 33–41.
16. Frazer, R. A., Duncan, W. J., and Collar, A. R., *Elementary Matrices,* New York: Macmillan, 1946.